(12) UK Patent Application (19) GB (11) 2 287 819 (13) A

(43) Date of A Publication 27.09.1995

(21) Application No 9505689.1

(22) Date of Filing 21.03.1995

(30) Priority Data
(31) S940265 (32) 25.03.1994 (33) IE

(71) Applicant(s)
GIST Limited

(Incorporated in Ireland)

Unit 25, Sandyford Office Park, Sandypark, Dublin 18,
Ireland

(72) Inventor(s)
Kenneth Blowers
Joseph Corcoran
Katherine Crean

(74) Agent and/or Address for Service
J A Kemp & Co
14 South Square, Gray's Inn, LONDON, WC1R 5LX,
United Kingdom

(51) INT CL^6
G06K 9/36

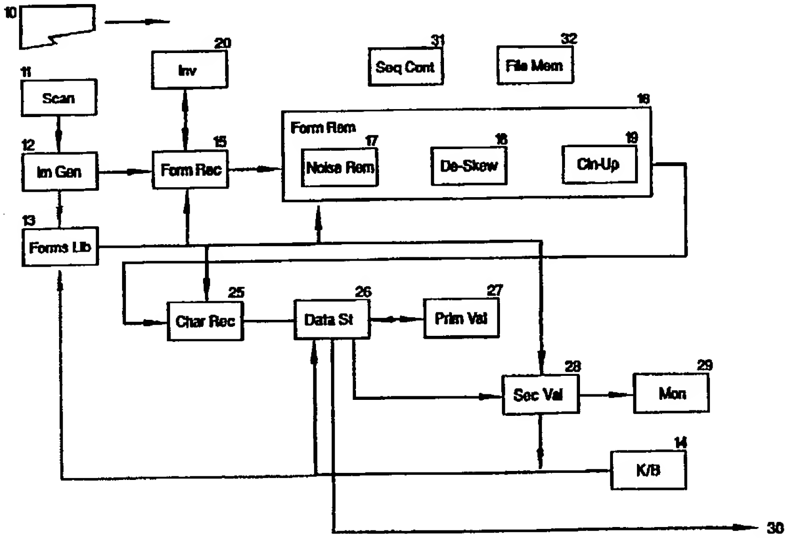
(52) UK CL (Edition N)
G4R REX RPN RPX R1A R1B R1C R1D R10F R10X R11D R11E R11F R3A R3C R3X R8A R8F R9B R9C
U1S S2123 S2210 S2273

(56) Documents Cited
US 5235654 A

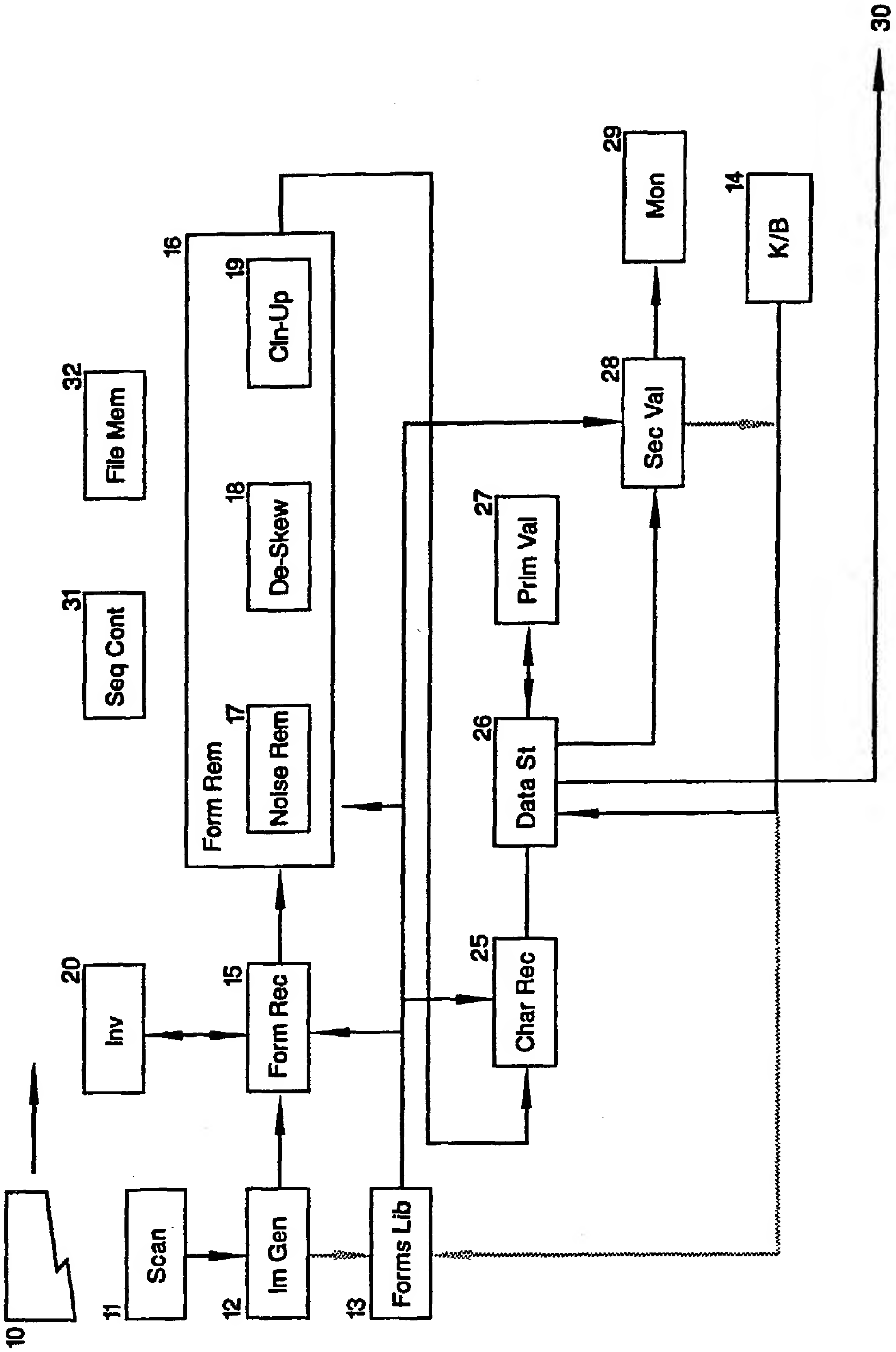
(58) Field of Search
UK CL (Edition N) G4R REX RHB RPA RPE RPN RPX
INT CL^6 G06K 9/00 9/03 9/20 9/32 9/36 9/54 9/60 9/78 9/80 17/00
Online:WPI

(54) Automated forms processing

(57) Vouchers (such as credit card vouchers) normally have a fixed layout, with areas on which fixed information is printed and a plurality of zones in which information has been entered manually; a zone normally consists of several sub-zones, one per character. To read such vouchers, a scanner 11 and image generating means 12 scan the voucher and create a voucher image file. A forms library 13 stores a plurality of form type images, including images in both a correct and an inverted orientation for each form type. A form recognition unit 15 compares the voucher image with the form type images to identify the form type of the voucher. Form removal means 16 remove the fixed form information from the voucher image. A character recognition unit 25 identifies the character (if any) in each sub-zone. Validation means 27 and 28 then validate the information so read, using a monitor 29 and a keyboard 14 for operator control.



GB 2 287 819 A


10
11
Scan
12
Im Gen
13
Forms Lib
20
Inv
15
Form Rec
16
Form Rem
17
Noise Rem
18
De-Skew
19
Cln-Up
31
Seq Cont
32
File Mem
25
Char Rec
26
Data St
27
Prim Val
28
Sec Val
29
Mon
14
K/B
30

AUTOMATED FORMS PROCESSING

The present invention relates to the automatic processing of documents, and more specifically to the recognition of written information thereon.

In many financial institutions, documents are processed in bulk, and information has to be captured from the documents and entered into an automatic processing system. The documents are frequently pre-printed forms having defined zones in which information is entered often in a manner which is constrained to some extent, eg by having to be entered character by character in marked squares, but is entered manually. A typical example, which we will use from here on, is the processing of credit card vouchers. (We will use the term 'written' for information entered when the voucher is actually used to record a transaction; this information will often be entered manually, but may be entered by some form of mechanical printing device.)

(Obviously, the forms may also have certain regions in which information is printed or otherwise recorded in a form designed for automatic capture, eg as printed characters in a suitably stylized font, magnetic ink characters, or bar codes. The capture of this information is therefore relatively simple and this type of information will be ignored herein.)

Currently, information is generally captured from such vouchers manually, by operators who read the information from the forms and key it into an automatic system using keyboards. This capture process is expensive and requires manual checking when key errors do occur, usually as a result of operator lack of consentration.

It would be desirable to be able to read such information automatically. Many automatic character recognition techniques have been proposed over the years. Many of these, however, are unsuitable for the present task of voucher processing. Some are designed for recognizing characters which are specialized in some way, eg as being written in magnetic ink, or requiring the characters to be controlled more precisely than those typically found on such vouchers. It has not so far proved practicable to perform automatic reading of information from vouchers and the like.

The general object of the present invention is to provide an improved system for automatically reading vouchers and the like.

Accordingly the present invention provides apparatus for reading vouchers or the like, each voucher consisting of areas on which fixed information is printed and a plurality of zones in which information has been written, the apparatus comprising:

scanning means for scanning a voucher to create a voucher image file;

means for storing a plurality of form type images, including images in both a correct and an inverted orientation for each form type;

form type identification means for comparing the voucher image with the form type images to identify the form type of the voucher;

inverter means for inverting voucher images identified as matching an inverted form type;

form removal means for removing the fixed form information from the voucher; and

character recognition means for recognising the character (if any) in each zone.

Each zone is preferably divided into one or more subzones, with one character being written in each subzone. This enables the complexity of the character recognition means to be reduced, since those means will then only have to recognize individual characters, instead of having to identify the boundaries separating adjacent characters.

A voucher reading system embodying the invention will now be described, by way of example, with reference to the drawing, which is a block diagram of the system.

The system is divided into two main portions: a first portion in which the vouchers exist essentially as graphic images, and a second portion in which the vouchers exist essentially as decoded character information. The first portion is concerned broadly with manipulating the voucher image into a form in which the desired characters can be read from it; the second portion is concerned broadly with checking and validating the characters so read.

Before the system can be used to process actual vouchers, it must be suitably intialized. For convenience, we will use the terms "forms" or "form types" for the various types of voucher. The purpose of this initializing is to store, in a forms library unit 13, details of all possible forms (types of vouchers) which the system is capable of recognizing. Initializing can obviously be done in a variety of ways. If desired, the system itself can be used for initializing. For this, a sample of each type of voucher which the system is to be able to process is scanned by a scanner 11 which feeds an image generator 12. The connection from unit 12 to unit 13 is shown by a broken line, as this connection is used only for this initializing procedure.

The forms library will thus contain, after initialization, a set of standard forms or templates, which are then used by the system for the processing of actual vouchers. These templates contain the images of the forms, as scanned by the scanner 11, but also contain various further items of information, regarding the location of the areas or zones of the vouchers in which information is filled in by the users, the nature of the information which is to be entered in those zones, etc. These further items may be entered or generated by any suitable means, shown for convenience here as a keyboard 14. Commercially available systems, such as FORMOUT from TIS (Tele Information Systems), may be used for some parts of this process.

It may be convenient for each template to consist of a single file, with the appropriate items of information being taken from that file as required at the various stages of processing of the vouchers; it is preferred, however, for each template to consist of a set of files, each containing the items of information required for a different stage or group of stages of voucher processing.

Each form is entered into the forms library twice, by being scanned in its correct orientation and then scanned in the reversed or inverted orientation (ie rotated through 180°). This enables the system to recognize vouchers which are fed in the wrong way round.

(The system will not, of course, be able to recognize vouchers which are fed in back to front, ie with the printing on the side of the voucher away from the scanner.) As will be seen, if a voucher is reversed, it is inverted during its processing, and the template for that form type in its normal orientation is used for subsequent processing.

Turning now to the actual processing of vouchers, a voucher 10 is scanned by the scanner 11 which feeds the image generator 12. The scanner 11 is conventional, and the image generator 12 is also essentially conventional, generating an image of the voucher in a standard format. This format can conveniently be a TIF (Tag Image File) style format, which consists of the actual image (pixel) information, normally in compressed form, together with various parameters such as the DPI (dots per inch) ratio, the pixel dimensions, and an orientation code. The output of the image generator 12 is a file containing the image in the TIF format.

When the voucher has been scanned and its image formed by unit 12, the voucher image file is passed to a form recognition unit 15. This unit compares the image of the voucher with the templates of all forms in succession in the forms library 12, and determines which of the forms the voucher matches. A form type identifier is added to the voucher image file. (If the voucher does not match any of the stored form type templates, it cannot be processed by the system and must be processed manually.)

The voucher image file produced by the form recognition unit 15 is passed to a form removal unit 16. The template for the form (as defined by the form type identifier in the voucher image file) includes what is in effect a mask defining the zones in the form in which variable information is to be written. This template is extracted from the forms library 13, and the form removal unit deletes, from the image of the voucher, those parts which represent fixed portions of the form, leaving only the zones defined by the template mask. Commercially available systems, such as certain systems produced by TIS (Tele Information Systems), may be used for some parts of this process.

The image removal unit 16 includes noise removal, de-skewing, and clean-up functions. These are shown for convenience as separate units 17 to 19; however, the de-skewing in particular may be combined with the form recognition and/or removal, since the details of the form removal may require localized adjustment of the fitting of the form template to the voucher image to achieve optimum removal of the fixed portions of the form, and that fitting may involve slight rotation of the form template and/or voucher image. Also, the registration of the image can be adjusted if desired. Commercially available systems may be used for some parts of these processes, such as the digital filtering (noise removal and clean-up).

The noise removal unit 17 removes isolated spots from the voucher image: an isolated spot can conveniently be defined as an isolated group of 1, 2, or 3 black pixels. The de-skewing unit 18 adjusts the orientation of the image to compensate for possible slight skewing of the voucher image; such skewing may result from, eg. a physical skewing of the voucher itself as it is scanned by the scanner 11, slight creasing of the voucher, or skewed printing of the voucher. The clean-up unit 19 may be used to reduce fuzziness of borders in the image, join up broken lines in the image, and so on: in particular, it may be used to join up lines which crossed form lines which have been removed and so have been broken by the removal of those form lines.

If the form type of the form is an inverted form the voucher image file is then passed to an inverter unit 20, which inverts the form. Inverting the voucher image involves rotating the image by 180°. This rotation is achieved by mapping every pixel in the voucher image to a point that is the same horizontal and vertical distance from the bottom right corner as the originating point is from the top left corner. Mapping the points in this way ensures that there is no change in the dimensions of the voucher image during the inversion process.

During the various stages of processing it is convenient to maintain the voucher image in a compressed form. The voucher images are maintained in TIFF Group 3 format, which is an industry standard format for imaging. The TIFF Group 3 format uses a modified form of LZW compression to compress the voucher image. This algorithm developed by Lempell et al replaces recurrent patterns of bytes with tokens that represent that pattern. In this way groups of recurring bytes are replaced with a single byte, thus reducing the disk space necessary to store the image. It is convenient to maintain the voucher image (in its various stages through the processing) in a compressed form, so this inversion may involve slight adjustments to the size parameters of the voucher image file.

The voucher image file can then be returned to the forms recognition unit 15, as shown, for re-identification of its form type. Alternatively, the voucher image file could (after inversion) be passed direct to the form removal unit 16. In that case, the form type identifier in the voucher image file must also be changed to the corresponding non-inverted form type. If desired, this can be achieved by using a particular bit in the form type identifiers to distinguish between the inverted and non-inverted form types, with subsequent units masking off that bit when using the form type identifier.

As noted above, zones are normally defined on the original voucher by boxes which define the areas into which information may be written. A zone may be intended to receive a single character, but most zones are intended to receive a plurality of characters, and are divided into subzones, each of which is intended to have a single character written in it. The zones and subzones for the different form types are defined by the templates in the forms library 14.

The voucher image file is then passed to a character recognition unit 25. This unit attempts to identify the character in each subzone in the voucher image; the locations of the zones of the voucher image, and the subzones within the zones, are defined by the template for the form type. (The markings identifying the zones and subzones have of course been removed from the voucher image file by the forms removal unit 16.) The character recognition unit 25 produces, for each subzone, a character identification, together with a confidence level, the coordinates of the location of the subzone on the voucher image, and the size of the character. (If the character identification is uncertain, two or more possibilities may be given.) Commercially available systems, such as the NESTOR ICR engine from Nestor (US), may be used for some parts of this process.

The character recognition unit 25 is effectively the interface between the two main portions of the system, the graphical image processing portion and the portion in which the vouchers exist essentially as decoded character information. The unit 25 produces a voucher data file for the voucher, containing a list of zone and subzone contents in numerical and character code, eg ASCII form. This voucher data file is passed to a storage unit 26.

The voucher information, which is now in abstract form, is now subjected to a series of validation operations in a primary validation unit 27, which involves checking the individual characters. This unit checks that the character height is consistent with the character; thus a character which is read as an "o" but is of very small height is likely to be a decimal point. It also checks the character against a character set for that zone or subzone, as defined by the form type; typical character sets are numeric, alphanumeric, and possibly checkmark (for a box which may be checked or left empty, or may be checked with a tick or a cross). This primary validation may resolve some uncertainties; eg if the character may be "1", "l", or "L", the ambiguity is resolved and the character identified as "1" if the subzone character set is numeric. The result of such resolution is fed back to the voucher data file in storage unit 26.

The voucher information is preferably also subjected to secondary validation in a secondary validation unit 28. This unit uses a variety of semantic-type checking techniques, operating generally on the character strings which occupy complete zones. Among the techniques which can be employed in suitable situations are the following:

Custom dictionary. For some zones, a complete list (the dictionary) of all possible "words" (ie sequences of characters) is predefined; eg a list of currencies, country names, etc. The word in the zone can be checked against the dictionary.

Check digit. An account number may include a check digit, which can be checked for arithmetical consistency.

Calculated zones. The contents of one zone may be determined by the contents of other zones, eg the contents of one zone may be the sum of the contents of other zones. Such relationships can be used to perform checks for arithmetical or other consistency.

Database look-up. The contents of two zones on a voucher may be related, eg account number and name. The account number from one zone may thus be used to look up the name in a database, and the name for that account number checked against the name on the voucher.

For some of these checks, it may be desirable to use fuzzy matching. One convenient form of such matching is "3cc" or 3 consecutive character comparison. This involves working along the string of characters in a zone until 3 consecutive characters are found all with high confidence levels, searching the database for entries with this 3-character string and then checking the entries so found for a match with the entire zone. Fuzzy matching logic can also be used to match entries such as "Bloggs" and "Bloggs Ltd" (but not "Bloggs Overseas") with "Bloggs Limited".

If a check fails, then it may be possible to correct the fault character, either automatically (eg changing "Blogg Limited" to "Bloggs Limited", where the database contains nothing else similar to "Bloggs Limited"), or manually, with the image of the field being displayed on the monitor 29 for the operator to read and to enter manual corrections via the keyboard 14.

If all the secondary validation checks eventually succeed, then the voucher data file has been successfully validated, and can be passed to some further apparatus over line 30. The character size and coordinate information will of course be deleted from the file for this.

If secondary validation is not possible, then the confidence level of characters which are below an acceptable level is manually validated. For this, the coordinates of the character are used to select the character from the voucher image file (preferably in the form in which it was processed by the character recognition unit 25) and displayed on a monitor 29. An operator then enters the character identification into the keyboard 14, and the contents of the voucher data file is updated accordingly. The system may be arranged to limit such manual amendment of a character in the voucher data file to "similar" characters, so that a numeric character "3" could be amended to say "8" but not to "4".

The voucher data file is then re-submitted for secondary validation where only after failure where no characters remain whose confidence level is below the acceptable level, is the voucher image file processed manually. This reduces the amount of manual validation required as often check digits and other means can indicate whether a character is correct or not without requiring manual input from an operator.

We have assumed so far that the voucher exists, in the system, as just two files, a voucher image file and a voucher data file, which undergo a variety of modifications as the processing proceeds. These two files can conveniently have the same filename (which will normally be an arbitrary identifier) but different filetypes. However, it may be convenient for the image file to have its name and/or filetype modified as it undergoes the various stages of processing, so that a set of files is created representing the various processing stages. (The voucher data file may be similarly treated.)

If desired, copies of the file or files at suitable stages can be stored in archive storage. The initial image file from the image generator unit 12 will of course contain more information than the file as presented to the character recognition unit 25, but the size of the file will typically be reduced several times by the processing, so storage of its final version may be preferable if large numbers of files have to be stored for long periods.

The description so far has been in terms of the processing of a single voucher. In practice, of course, the system will have to process a large number of vouchers. To control this, there is a sequence control unit 31 which contains a record for each voucher, and a file memory unit 32. As a voucher is processed, so the files created for it are stored in the file memory 32. (The data store 26 can conveniently be part of the file store 32, even though it is shown separately.) The voucher record in the sequence control unit 31 for each voucher can conveniently contain the voucher identifier (as described above), pointers to the locations in the file memory 32 of the various files associated with the voucher, and a series of fields (which can generally be single characters) indicating the progress of the processing of the voucher through the various stages of processing (eg pending, being performed, and completed for most of the stages). Also, certain control entries which have been described above as being made in the voucher image (or data) files may actually be made in the voucher record. The sequence control unit can conveniently be included in a master database which also contains the various voucher image and text files.

This organization of the sequence control unit and the file memory allows the system to be implemented by a group of processors of the PC type, with various units capable of performing various operations. As a unit becomes free, so it can check the records in the sequence control unit to find a processing operation waiting to be performed.

It will be realised that the scanner 11 and image generator 12 may be incorporated into a fax machine which receives voucher images over a phone line and subsequently generates the voucher image file. Some pre-processing of the faxed voucher image file may then be required to convert its format into a format capable of being recognised by the remainder of the system.

CLAIMS:

1. Apparatus for reading vouchers or the like, each voucher consisting of areas on which fixed information is printed and a plurality of zones in which information has been written, the apparatus comprising:

scanning means for scanning a voucher to create a voucher image file;

means for storing a plurality of form type images, including images in both a correct and an inverted orientation for each form type;

form type identification means for comparing the voucher image with the form type images to identify the form type of the voucher;

inverter means for inverting voucher images identified as matching an inverted form type;

form removal means for removing the fixed form information from the voucher; and

character recognition means for recognising the character (if any) in each zone.

2. Apparatus according to Claim 1, including means for returning said inverted voucher images to the form type identification means for re-identification.

3. Apparatus according to either previous claim, including means for generating form type images by scanning sample vouchers.

4. Apparatus according to any previous claim, including

means for processing the image to improve its quality between the form removal means and the character recognition means.

5. Apparatus according to any previous claim, wherein the validation means includes means for checking individual characters against character sets associated with the zones of those characters.

6. Apparatus according to Claim 5, wherein each zone comprises one or more subzones in which information has been written one character per subzone, and the validation means includes means for checking individual characters against character sets associated with the subzones of those characters.

7. Apparatus according to any previous claim, including validation means including operator controlled means, for validating the information so read.

8. Apparatus according to Claim 7, wherein the validation means includes means for checking the consistency of the character strings in complete zones against criteria which the contents of those zones should satisfy.

9. Apparatus for reading vouchers or the like, substantially as herein described and illustrated.

Patents Act 1977 **:xaminer's report to the Comptroller under Section 17** **(The Search report)**	Application number GB 9505689.1
Relevant Technical Fields (i) UK Cl (Ed.N) G4R (REX, RHB,RPA, RPE, RPN, RPX) (ii) Int Cl (Ed.6) G06K 9/00, 9/03, 9/20, 9/32, 9/36, 9/54, 9/60, 9/78,9/80, 17/00	Search Examiner J DONALDSON Date of completion of Search 25 APRIL 1995
Databases (see below) (i) UK Patent Office collections of GB, EP, WO and US patent specifications. (ii) ONLINE: WPI	Documents considered relevant following a search in respect of Claims :- 1 TO 9

Categories of documents

X:	Document indicating lack of novelty or of inventive step.
Y:	Document indicating lack of inventive step if combined with one or more other documents of the same category.
A:	Document indicating technological background and/or state of the art.
P:	Document published on or after the declared priority date but before the filing date of the present application.
E:	Patent document published on or after, but with priority date earlier than, the filing date of the present application.
&:	Member of the same patent family; corresponding document.

Category	Identity of document and relevant passages	Relevant to claim(s)
A	US 5235654 (ANDERSON) see column 6, lines 44-56, column 22, lines 36-39	

Databases:The UK Patent Office database comprises classified collections of GB, EP, WO and US patent specifications as outlined periodically in the Official Journa (Patents). The on-line databases considered for search are also listed periodically in the Official Journal (Patents).